DOCUMENT RESUME

ED 395 952 TM 025 020

AUTHOR Arnold, Margery E.
TITLE The Effects of Two Types of Sampling Error on Common Statistical Analyses.
PUB DATE 25 Jan 96
NOTE 19p.; Paper presented at the Annual Meeting of the Southwestern Educational Research Association (New Orleans, LA, January 25, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Error of Measurement; Research Methodology; *Sample Size; Sampling; *Statistical Analysis; Statistical Distributions
IDENTIFIERS *Sampling Error; *Variability

ABSTRACT

Sampling error refers to variability that is unique to the sample. If the sample is the entire population, then there is no sampling error. A related point is that sampling error is a function of sample size, as a hypothetical example illustrates. As the sample statistics more and more closely approximate the population parameters, the sampling distributions have less and less variability, and the standard error (standard deviation of the sampling distribution) decreases until there is no sampling error. Sampling error is two dimensional in that the effects of sampling error increase and decrease as a function of how many of the population members are sampled as well as who is sampled. In a one dimensional framework, sampling error is simply the difference between the results one would have obtained if one sampled the entire population and the results one did obtain. An appendix presents a computer program to sample scores randomly. (Contains three figures, two tables, and eight references.) (SLD)

Rev. 1/21/96

The Effects of Two Types of Sampling Error on Common Statistical Analyses

Margery E. Arnold

Texas A&M University

Paper presented at the annual meeting of the Southwestern Educational Research Association, New Orleans, LA, January 25, 1996.

Many a researcher has been overheard to say " the differences between the control group and the treatment group are too different to be attributable to 'chance'." When asked to explain what they mean by this, or how they know that this is true, researchers are often unable to give an explanation. The present paper explains the "chance" to which the researcher is referring, as well as why this 'chance' occurs and how it affects common statistical analyses such as t-tests, ANOVA and stepwise multiple regression.

To study a phenomenon, a researcher must define his or her population and then observe, measure or otherwise quantify this phenomenon in that population. Usually it is not possible to measure every member of that population, therefore the researcher must rely on studying the phenomenon in a portion of the population, called a sample. This, of course, creates a problem.

> No matter how good a sampling method is used to draw a sample, it is clear that a sample can never reproduce exactly the various characteristics of the population unless the population itself is taken as the sample and a census is carried out. The resulting discrepancies between the sample estimates, and the population values that would be obtained by enumerating all the units in the population in the same manner in which the sample is enumerated, are termed *sampling errors*. (Sukhatme, Sukhatme, Sukhatme, & Asok, 1984, p. 2)

In other words, "sampling error" refers to variability that is unique to the sample. If the sample *is* the population, then there is no sampling error. As stated earlier, however, it is usually not possible to take a census. A related point is that sampling error is a function of sample size. As the sample size decreases from the size of the population, sampling error is likely to increas. As Thompson (1994) explains, "Samples (and thus the statistics calculated for

them) will potentially be less representative of populations as sample sizes are smaller." (p. 5). An example illustrates this.

Suppose for example, a researcher wants to know levels of depression in the 100 people who lived in Hypothetical Village, TX in 1994. The researcher knows that all of the inhabitants (the population) are alive and that they no longer live in Hypothetical Village, TX, because there was a nuclear accident there at the end of 1994. Suppose the researcher has resources to find and include in her sample only 50 persons in the population. What might happen if she can only sample 10? Will the sampling error increase? To make the example more clear Table 1 provides parameters (mean, standard deviation, mode, median, maximum and minimum of the distribution of the 100 scores) that have been calculated using the depression scores of all 100 inhabitants. Also in the table are 8 hypothetical samples of the 100 persons in the population. In the first four samples, the sample size is 50. In the next four samples, the sample size is 10. (These statistics and parameters were obtained using SPSS. The program used to generate these results can be found in Appendix A.)

Sampling Error and Sampling Distributions as Affected by Sample Size.

To make the point of how sampling error is effected by sample size clear, this discussion will focus only on means. (The other statistics are provided for further exploration of the effect of sampling error.) Note that the range of the means of the samples of 50--samples #1, 2, 3 & 4-- is 3.16 (83.52-80.36 = 3.16), whereas the range of means for the samples of 10--samples #5, 6, 7, & 8--is 5.3 (82.50 - 77.20 = 5.3). There is more variability, as measured by range of sample means, in the samples of smaller size (n=10) than in the samples of larger size (n=50). There is also more deviation from the mean of the population (81.25) in the smaller samples (77.20, 77.20, 82.50, 74.10) than in

the larger samples (81.34, 83.52, 81.66, and 80.36). This deviation of the sample statistic from the population parameter is sampling error. Thus, sampling error is usually greater in smaller samples as opposed to larger samples.

The word "usually" is used because it is *possible* for a sample statistic from a small sample to be more accurate (closer to the population parameter) than a sample statistic from a larger sample. However, this is not very *probable*. Probability (i.e., "chance") enters into this discussion in the form of underlying distributions. An underlying distribution represents all possible outcomes of a particular event. In this case, the event is a sample statistic--the mean. Underlying distributions that are used to estimate the probability of an obtained sample statistic are called sampling distributions. Sampling distributions are defined by two different constraints: a population parameter and sample size. Figure 1 illustrates two different sampling distributions, one for n=10 and one for n=50. In this example there are two different sample sizes, so there are two different sampling distributions (one for n=50 and one for n=10). There is, however, only one population parameter, as both of the sample types are drawn from the same population, which has a known mean of 81.25. (Often population parameters are assumed, not known. This subject will be discussed later.)

When examining the distributions in Figure 1, note that it is possible for sample of 10 to contain a mean very close to the true mean in the population. The probability of this happening, however, is not as high as it is in the sample of size=50. The reason for this is that the sampling distribution of means for a smaller sample size has much more sampling error as indicated by variability. In the sampling distribution for n=10, 99.74% of the scores range from 47.86 to 114.64, whereas in the samping distribution of means for samples containing 50 members 99.74% of the scores range from

74.56 to 87.94. Wider variability affects probability in that it spreads probability out over a wider range of numbers; therefore, every possible mean contained between -3SD and +3SD in the sampling distribution of n=10, has a lesser chance of occuring in the long run than each of the possible means contained between -3SD and +3SD in the sampling distribution of n=50.

For example, notice the mean scores at SD=-1. If the mean is 81.25 in the population, there is an equal likelihood that one will obtain $\underline{M}$=79.02 (n=50) as there is that one will obtain $\underline{M}$=70.12 (n=10). Said differently, the 34% of probability that exists between the mean and -1SD is spread out for 2.23 units (81.25-79.02=2.23; n=50) or is spread out for 11.13 units (81.25-70.12=11.13; n=10). As illustrated by this example, it is more likely that there will be less sampling error in samples of larger sizes than in samples of smaller sizes; 79.02 (n=50) more closely approximates the population value (81.25) than does 70.12 (n=10).

Hinkle, Wiersma and Jurs (1994, p. 154) illustrate clearly the relationship between sample size and variability in sampling distributions. An adaptation of their illustration can be found in Figure 2. Notice that the first drawing (A) is a frequency distribution of the population. In other words, the verical axis represents the amount of times that a value occurs in the population and the horizontal axis contains the values that occur in the population. (It may be easier to think of the scores as something specific such as SAT verbal scores.) The subsequent drawings (B-F) are illustrations of *distributions of sample means*, each sample came from the population in drawing A. Note that the mean of each of the distributions (ie., the mean of the means) is the same as the mean for the population, μ = 500.

The only difference between each of the sampling distributions in drawings B through G is sample size. Drawing B depicts the distribution of

means for infinite numbers of samples of size n=1. Note that the distribution in B, a distribution of sample means, is exactly identical to the distribution in A, a distribution of the population. Because the sample size is n=1 for the samples depicted in drawing B, each sample has only one member and the value assigned to that member *is* the mean for that sample. In the sampling distribution depicted in drawing C each of the samples contains two members. Notice that there is less variability in sampling distribution C than there is in sampling distribution B. Another way of saying 'there is less variability" is to say "the frequency distribution curve is higher and less wide" or to say "the standard deviation of the sampling distribution (also known as the standard error) is decreased." The distribution is "higher" because there are more sample means that exactly equal the mean of the sampling distribution (and also the mean in the population) in distribution C than there were in sample B. This trend of sample statistics being more and more likely to equal or more closely approximate the population parameter continues throughout the drawings as sample size increases. That is to say, that as more and more persons from the population are included in the sample, the sample statistics are more and more likely to equal or closely approximate the population parameter.

As the sample statistics more and more closely approximate the population parameters, the sampling distributions have less and less variability and the standard error (standard deviation of the sampling distribution) decreases until there is no sampling error. This is illustrated in drawing G of Figure 2. When samples of size infinity are drawn from the population, the mean of the sample exactly equals the population parameter for every sample of size n= infinity that is drawn. There is absolutely *no* variability in the resulting sampling distribution. The standard error

(standard deviation of the sampling distribution) equals zero; and there is absolutely *no* sampling error. Of course, it is impossible to calculate a mean for an infinite number of values. Nonetheless, the example explains what could theoretically happen if it were possible to have absolutely no sampling error.

Extending the example case: t-tests, and p values

The above example illustrates the effect of sampling error on descriptive statistics. Sampling error, of course, affects other analyses as well. Many researchers in education and psychology study the differences between a treatment group and a control group. To extend the Hypothetical Town example, a researcher might decide to study the effectiveness of a treatment for depression by using that treatment with a sample of persons from Hypothetical Town and comparing the post-treatment depression scores of the treatment group to the scores of another sample of persons from the same town who did not receive the treatment (control). Using a t-test or ANOVA to compare the mean of the depression scores of the treatment group ($\underline{M}_t$) to the mean of the depression scores of the control group ($\underline{M}_c$), is one way to study the effects of the treatment. Many researchers believe that if the difference between $\underline{M}_t$ and $\underline{M}_c$ is statistically significant (i.e., $\underline{p} < .05$ or $\underline{p} < .01$), then they can say that their treatment was effective.

Some researchers (See Carver, 1993; Shaver, 1993; and Thompson, 1993) would argue, however, that a $\underline{p}$ value does not indicate whether or not a treatment is effective. To illustrate their points that are relevent to sampling error, consider again the concept of a sampling distribution as affected by sampling error. Sampling distributions are used to derive $\underline{p}$ values. "$\underline{p}$ " has been variously defined. Shaver (1993) defines $\underline{p}$ as,

> the probablility of a result occuring by chance in the long run under the null hypotheseis with random sampling and size n; it provides no conclusion about the probability that a particular result is attributable to chance. (p. 300)

Thompson (1994) defines p as that which answers the question,

> Assuming the sample data came from a population in which the null hypothesis is (exactly) true, what is the probability of obtaining the sample statistics one got for one's sample data with the given sample size(s)? (p. 5)

From these definitions it is clear that sampling distributions (and thus p values) are based upon two criteria: population parameter (in this case, mean) and sample size. The researcher knows the sample size for the two samples that he or she has drawn, treated and tested (in this case, either 10 or 50). The researcher does not know, however, the true mean in the population. Therefore, he or she must make an assumption upon which to base the sampling distribution.

For example, suppose that M_t=74 and M_c=82. Because the researcher does not know what the true population mean is, he or she might construct a sampling distribution based on the null hypothesis. The null hypothesis in this case would be that there is no difference between M_t and M_c or (M_t-M_c=0). Using Figure 3, if our sample size=50 (lower amount of sampling error), and M_t-M_c=74-82= 8, then $p < .01$. If, however, our sample size=10 (greater amount of sampling error), and M_t-M_c=74-82= 8, then $p < .16$, which by the standards of most researchers, is not statistically significant. Even though the effect size is the same--M_t-M_c=74-82= 8-- the results of the larger sample are statistically significant, while the results of the smaller sample are not. Statistical significance, then, is a function of sampling error as created by

sample size, as much as it is a function of the magnitude of the effect size. In fact the calculated test statistic t, literally is the difference in the two means divided by the standard deviation of the sampling distribution associated with a given null hypothesis and a given sample size (called the standard error).

Two Dimensional Sampling Error

Thompson (1995) offers a definition of sampling error that differs from the Sukhatme et al. (1984) definition. "Sampling error is variability in sample data that is unique to the given sample, and therefore cannot be reproduced in subsequent samples." The definition put forth by Thompson (1995) seems to refer to the two-dimensional nature of sampling error. Error in this definition refers to variance, an area concept, as opposed to the one-dimensional phenomenon described by Sukhatme et al. (1984): population parameter-- sampling statistic= sampling error.

A two dimensional concept of sampling error can be illustrated by referring to Table 1 and comparing sample #5 and sample #6, both of sample size n=10. According to the definition given by Sukhatme et al., (1984), it appears that sample #5 and sample #6 have the same amount of sampling error. (Both have a mean of 77.20 and therefore, both sample means differ from the population mean of 81.25, by the same amount, 4.05 units.) Using the Thompson (1995) definition and looking at variances, however, one notices that the possibility exists that sample #6 contains more sampling error than sample #5. The individual membership of sample #5 and sample#6 are enumerated in Table 2. Notice that even though the means are the same (77.20), there is a difference in variability as measured by variance. Table 2 illustrates how these two samples come to have differing variances. The variance of sample #5= 99.51, whereas the variance of sample #6= 135.51.

How can this be? The difference in variability between sample #5 and sample #6 is a function of *whom* is sampled. Both samples were randomly chosen, but sample #6 contains members with more extreme scores than does sample #5.

This definition of sampling error is important to understand because many statistical analyses (such as ANOVA and multiple regression) use a shared variance type of effect size. Sample #6 has more variance to share than does sample #5, simply because of *whom* is in the sample. It is possible therefore, that sample #6 could yield a higher effect size than sample #5 simply because of sampling error as defined by Thompson (1995). Unless the exact same people are sampled every time, it is highly unlikely that findings will be replicated exactly. This principle becomes even more important in analyses which make decisions based on variance accounted for, such as stepwise multiple regression.

> At a given step, the determination of which single variable to enter will enter variable X_1 over variables X_2, X_3 and X_4, even if X_1 is only infinitesimally superior to the other three variables. It is entirely possible that this infinitesimal advantage of variable X_1 over another variable is sampling error, given that the competitive advantage of X_1 is so small. (Thompson, 1995, p. 532)

Conclusions

As has been demonstrated sampling error affects the results of common statistical analyses. The effects of sampling error increase and decrease as a function of *how many* of the population members are sampled as well as *whom* is sampled. Sampling error can be conceptualized one-dimensionally as well as two-dimensionally. In a one-dimensional framework, sampling error is simply the difference between the results one

would have obtained if one would have sampled the entire population and the results that one did obtain, sampling only a portion of the sample. In a two-dimensional framework, sampling error can be described as variance created uniquely by each member of the sample. Both the one-dimensional and the two dimensional conceptualizations of sampling error are affected by *whom* is sampled and *how many* are sampled.

References

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61 (4), 287-292.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). Applied statistics for the behavioral sciences (3rd ed.). Boston: Houghton Mifflin Company.

Kirby, J. H., Culp, W. H. & Kirby, J. (1973). Manual for users of standardized Tests. Bensenville, IL: Scholastic Testing Service, Inc.

Shaver, J. P. (1993). What statistical significance testing is and what it is not. Journal of Experimental Education, 61 (4), 293-316.

Sukhatme, P. V. , Sukhatme, B. V., Sukhatme, S. & Asok, C. (1984). Sampling theory of surveys with applications. Indian Society of Agricultural Statistics.

Thompson, B. (1987, April). The use (and misuse) of statistical significance testing: some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868)

Thompson, B. (1994). The concept of statistical significance testing. Measurement Update, 4(1), 5-6.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement, 55, 525-534.

Table 1

Population parameters of the 100 people living in Hypothetical Town, TX and sample statistics for eight samples of differing sample size

	Pop.	Samples (n=50)				Samples (n=10)			
	N=100	#1	#2	#3	#4	#5	#6	#7	#8
Mean	81.25	81.34	83.52	81.66	80.36	77.20	77.20	82.50	74.10
Median	83.50	82.50	85.00	85.00	83.00	75.00	80.00	83.00	72.00
Mode	85.00	85.00	85.00	85.00	90.00	63.00	80.00	68.00	63.00
SD	10.55	9.59	8.83	11.00	11.25	9.97	11.64	8.53	10.98
Minimum	55.00	56.00	64.00	55.00	55.00	63.00	60.00	68.00	63.00
Maximum	99.00	99.00	99.00	99.00	99.00	95.00	97.00	95.00	97.00

Table 2

Demonstration of 2-dimensional Sampling Error as it occurs in two randomly chosen samples of n = 10

Sample #5 from Table 1

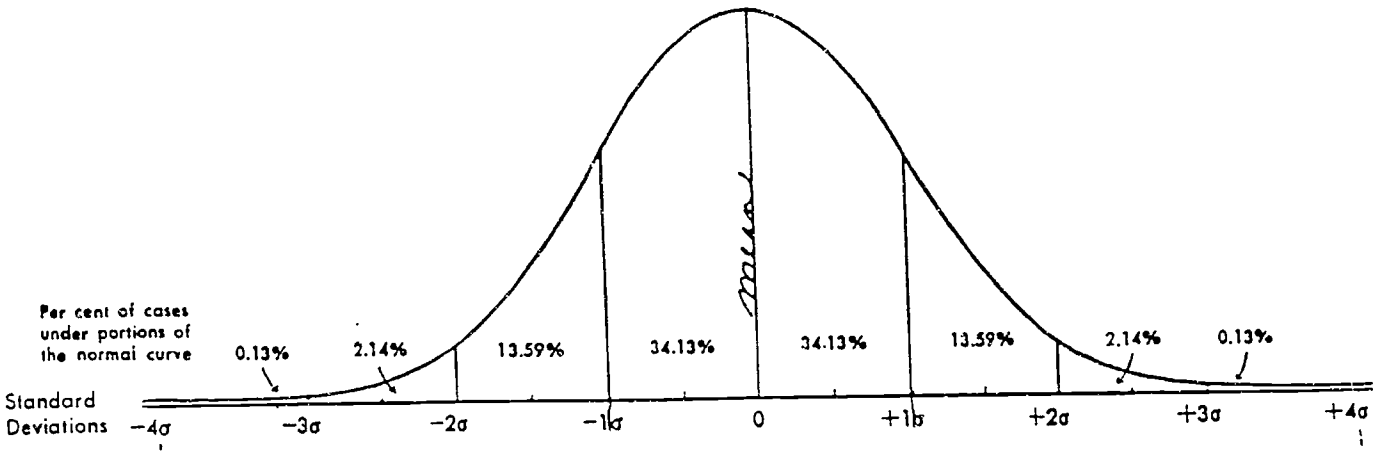
ID #	Depression Score	$X-x_i$	$(X-x_i)^2$	
97	63	14.2	201.64	
25	67	10.2	104.04	
10	70	7.2	51.84	
9	73	4.2	17.64	
14	74	3.2	10.24	
75	76	1.2	1.44	
68	81	-3.8	14.44	
45	85	-7.8	60.84	
42	88	-10.8	116.64	
86	95	-17.8	316.84	
	Mean=77.2		Sum=895.6	SD^2=99.51

Sample #6 from Table 1

ID #	Depression Score	$X-x_i$	$(X-x_i)^2$	
67	60	17.2	295.84	
99	63	14.2	201.64	
19	67	10.2	104.04	
100	71	6.2	38.44	
50	80	-2.8	7.84	
11	80	-2.8	7.84	
69	83	-5.8	33.64	
53	85	-7.8	60.84	
1	86	-8.8	77.44	
80	97	-19.8	392.04	
	Mean = 77.2		sum=1219.6	SD^2=135.51

Figure 1

The sampling distributions for means (n=50, n=10, population mean= 81.25).



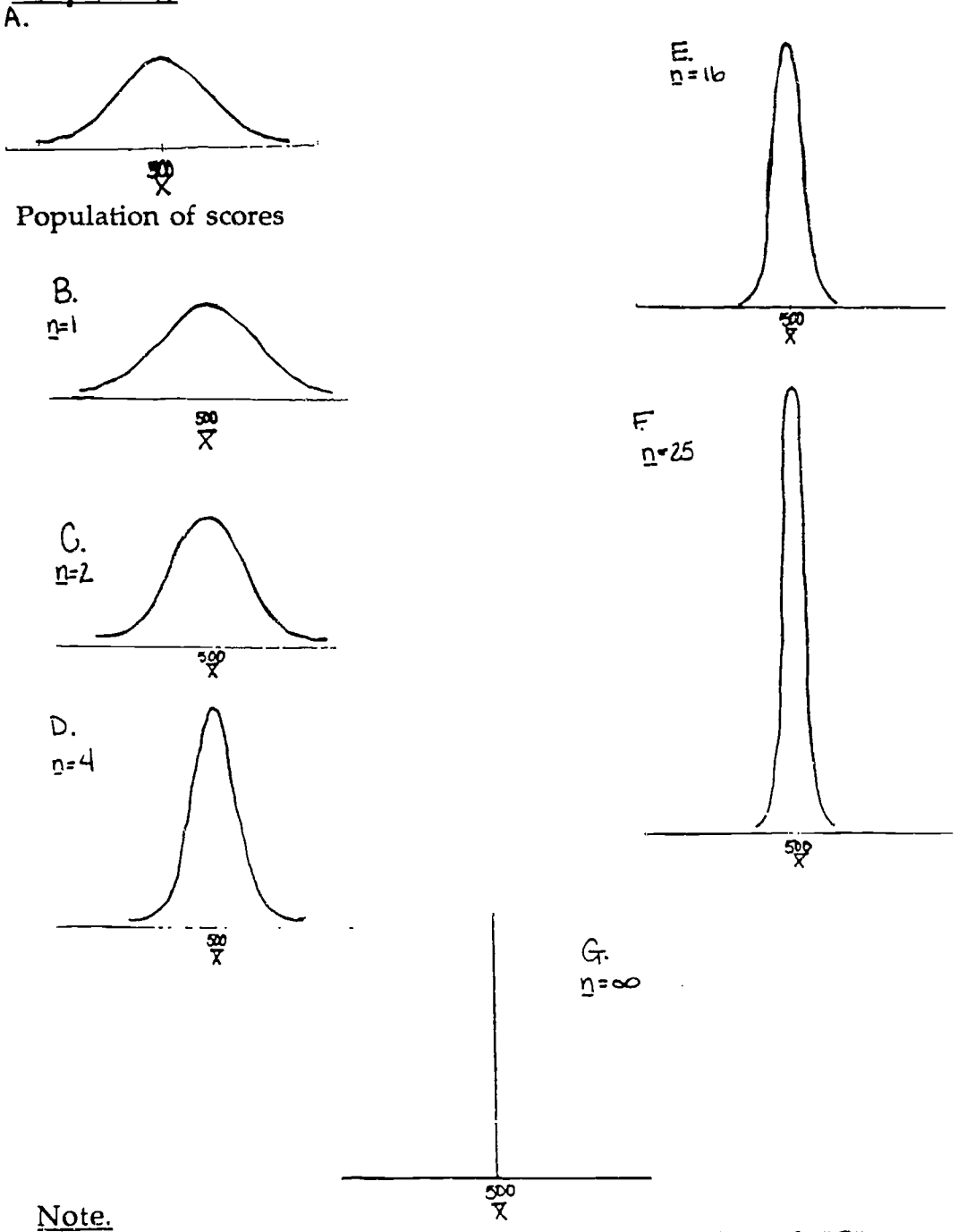
	SD=-3	SD=-2	SD=-1	Mean	SD=+1	SD=+2	SD=+3
n=50	74.56	76.79	79.02	81.25	83.48	85.71	87.94
n=10	47.86	58.99	70.12	81.25	92.38	103.51	114.64

Note. The scaling for the distributions for n=50 and n=10 are different. This was done so that points along the normal curve of the distribution could be highlighted and compared. If the two distributions were to be drawn separately, with equal scaling, the n=10 distribution would be flatter and wider than the n=50 distribution.

*Normal curve copied from Kirby, Culp & Kirby (1973).

Figure 2

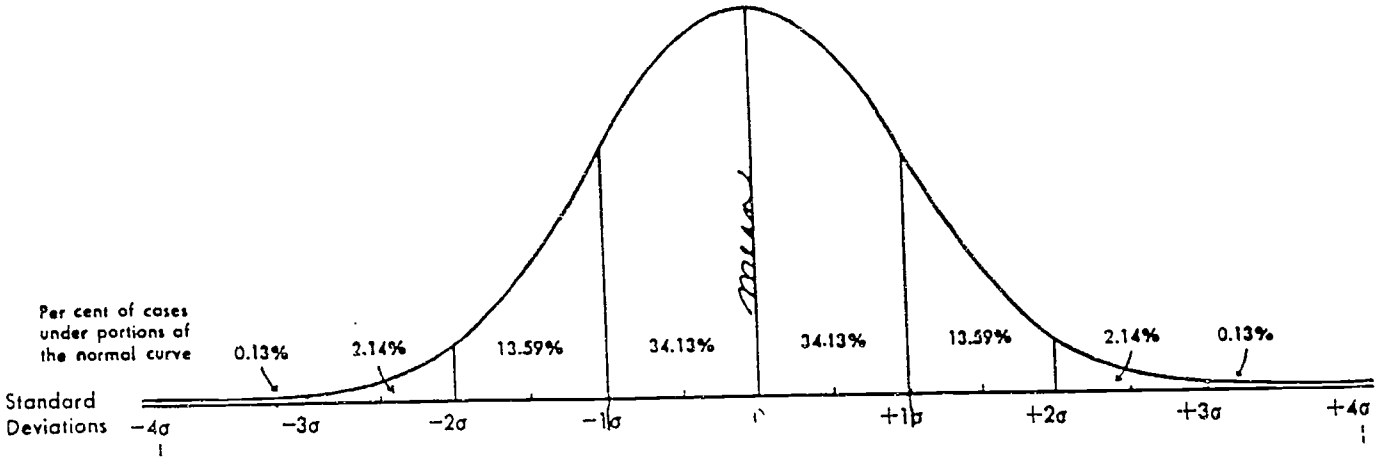
Normal distribution of a population with sampling distributions of differing sample sizes



Note.
Adapted from Hinkle, Wiersma & Jurs (1994). "B" through "G" are sampling distributions presenting distributions of parameter estimates and not of scores.

Figure 3

Sampling distribution for the differences between means for two sample sizes



	SD=-3	SD=-2	SD=-1	Mean	SD=+1	SD=+2	SD=+3
n=50	-5.61	-3.69	-1.92	0	1.92	3.69	5.61
n=10	-13.38	-8.92	-4.46	0	4.46	8.92	13.38

Note. The scaling for the distributions for n=50 and n=10 are different. This was done so that points along the normal curve of the distribution could be highlighted and compared. If the two distributions were to be drawn separately, with equal scaling, the n=10 distribution would be flatter and wider than the n=50 distribution.

*Normal curve copied from Kirby, Culp & Kirby (1973).

Appendix A

SPSS Program to randomly sample 10 and 50 scores from a population of 100

```
Data list file=ABC/ ID 1-3 DEPRESS 5-6
Frequencies variables= DEPRESS
 /HISTOGRAM
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '1: 1ST SAMPLE OF 50 FROM 100'
TEMPORARY
SAMPLE 50 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '2: 2ND SAMPLE OF 50 FROM 100'
TEMPORARY
SAMPLE 50 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '3: 3RD SAMPLE OF 50 FROM 100'
TEMPORARY
SAMPLE 50 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '4: 4TH SAMPLE OF 50 FROM 100'
TEMPORARY
SAMPLE 50 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '5: 1ST SAMPLE OF 10 FROM 100'
TEMPORARY
SAMPLE 10 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '6: 2ND SAMPLE OF 10 FROM 100'
TEMPORARY
SAMPLE 10 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '7: 3RD SAMPLE OF 10 FROM 100'
TEMPORARY
SAMPLE 10 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
SUBTITLE '8: 4TH SAMPLE OF 10 FROM 100'
TEMPORARY
SAMPLE 10 FROM 100
FREQUENCIES VARIABLES= DEPRESS
 /STATISTICS= DEF MODE MEDIAN
```